Sistema de reconocimiento multilenguaje del habla

Ali Montiel, Mario De Jesús, Raúl Hernández, Rubén Maldonado, Veronica Olvera, Yanette Morales y Leticia Flores-Pulido

Universidad Autónoma de Tlaxcala, Facultad de Ingeniería y Tecnología, Apizaco, Tlaxcala, México

{ilusion119,idmariodjc,chanbrawl,rumalo1791,vero.pink15,yanette_morales_salado,aicitel.flores}@gmail.com

Resumen. Este trabajo se comienza con la presentación de una serie de artículos relacionados con el Reconocimiento Automático del Habla. Se realiza un análisis de cada uno de ellos donde se obtienen datos relevantes y los que serán de gran ayuda para desarrollar la propuesta multilenguaje de un sistema de reconocimiento del habla aquí descrito. Existen varias técnicas que son aplicadas para lograr una efectividad más alta de los sistemas basados en Reconocimiento Automático del Habla. Entre las más utilizadas se encuentran los coeficientes cepstrales de Mel, el modelo oculto de Markov y Coeficientes Predictivos Lineales. Cada uno de los trabajos relacionados con el reconocimiento automático del habla presenta su propio modelo de lenguaje y un modelo acústico que permite tener un amplio porcentaje de efectividad. Las técnicas anteriormente mencionadas forman parte de la extracción de características de la prupuesta multilenguaje. El objetivo entonces es una propuesta de implementación que pueda reconocer diferentes clases de idiomas basado en una extracción de características bajo la combinación de técnicas como son los modelos ocultos de markov y los coeficientes de predicción lineal. En éste trabajo se muestra la etapa de extracción de formantes de tres corpus del habla de diferentes idiomas: PRESEEA, EUSTACE y DIMEX100.

Palabras clave: coeficientes de Mel, espectrograma, frecuencia, coeficientes predictivos lineales, modelo oculto de Markov.

1. Introducción

El proceso de reconocimiento automático del habla (RAH) dota a las máquinas de la capacidad de recibir mensajes orales. El reconocimiento automático del habla proporciona una nueva forma de interactuar con un computador, en este caso a través de la voz, este tipo de interfaces también son llamadas de usuario de voz, e interfaces basadas en el habla. Las tecnologías del habla son muy utilizadas en las aplicaciones de servicios telefónicos ofrecidos a los

usuarios para la realización de alguna operación bancaria. Las tecnologías del reconocimiento de voz se realizan bajo tres pilares, diseño de IVR/SIU, las ciencias del servicio y los factores humanos. Este tipo de tecnologías abre un gran abanico de aplicaciones prácticas como por ejemplo: (a) Sistemas de dictado, donde lo que se pretende es una transcripción textual lo más exacta posible de aquello que ha dicho un locutor. Y (b) Sistemas de diálogo, donde el objetivo es conceptualizar aquello que se ha captado por el sensor auditivo e inferir una respuesta. En definitivo, el reconocimiento automático del habla es un campo con gran interés práctico y que presenta problemas no precisamente triviales de resolver. Es por ello que se propone un sistema que reuna tres tipos de corpus del habla bajo diferentes idiomas: español de España, Inglés Británico y Español de México, que sea capaz de conformar tres clases de formantes que puedan ser discretizados por diferentes extractores de características y que además puedan ser reconocidos.

2. Estado del arte

En [1] se trabajó con un reconocedor que utilizó elementos independientes del contexto, denominadas "monófonos", como unidades básicas del modelo acústico. Para la creación de los modelos se emplearon modelos ocultos de Markov MOM de tres estados de izquierda a derecha del tipo semi-continuo asociados a cada uno de los 31 monófonos (30 fonemas + alófonos y un modelo de silencio). En [2] se presentan dos sistemas de análisis acústico del habla con aplicaciones a la descripción de segmentos de discurso espontáneo y un sistema de reconocimiento automático de habla espontánea orientado a la detección de palabras. En [3] se tiene como objetivo mejorar la interacción el hombre y la máquina, haciendo posible que un determinado dispositivo pueda rescatar información afectiva más que el contenido hablado por una persona. En [4] se menciona que el ruido de fondo está frecuentemente presente en ambientes donde se emplean sistemas de Reconocimiento Automático del Habla (RAH). Una señal ruidosa da lugar a una degradación en la tarea del reconocimiento debido al desajuste con el modelo acústico (MA). En [5] se plantea que la motivación principal es crear un sistema de reconocimiento automático del habla en el idioma español, el cual tiene como objetivo lograr altas tasas de reconocimiento en comparación con otros sistemas de su tipo. En [6] se considera también al ruido como uno de los principales factores a tener en cuenta en las aplicaciones reales del reconocimiento automático de voz. El rendimiento de los reconocedores se ve fuertemente afectado cuando la señal de voz es adquirida en un entorno ruidoso. En [7] se propone un algoritmo para el reconocimiento de personas en un canal telefónico. El algoritmo se basa en el comportamiento de las Redes Neuronales Artificiales (RNA), en particular, sobre el algoritmo Backpropagation. En [8] se presenta a Kaldi que es una herramienta que proporciona una biblioteca de módulos diseñados para acelerar la creación de sistemas automáticos de reconocimiento de voz para fines de investigación. Los efectos del modelado acústico y el conjunto de herramientas proporciona un marco para formantes bajo redes

neuronales mediante descenso de gradiente estocástico para el reconocimiento del habla.

3. Métodos de reconocimiento automático del habla

Existen varios métodos de reconocimiento del habla, los cuales no serán descritos a detalle, pero si serán mencionados a grandes rasgos para comprensión del lector.

3.1. Coeficientes predictivos lineales (Linear Predictive Coding)

Los CPL (coeficientes predictivos lineales) son un modelo para la producción de la señal de voz con la suposición inicial de que la señal de voz es producida bajo un modelo acústico muy especifico. Es un método para el modelado de la señal de voz y es de uso frecuente por los lingüistas como una herramienta de extracción de formantes. El análisis LPC es generalmente apropiado para modelar las vocales que son periódicas, salvo las vocales nasales. El LPC se basa en el modelo de fuente-filtro de la señal de voz.

El algoritmo consiste en lo siguiente:

- Pre énfasis: La señal de voz digitalizada, $s(n)$, se somete a un sistema digital de bajo orden, para espectralmente aplanar la señal y hacerla menos suceptible a efectos de presición finita posteriores en el procesamiento de la señal. La salida de la red de pre énfasis, está relacionada a la entrada de la red, $s(n)$, por la siguiete ecuación:

$$\tilde{s}(n) = s(n) - \tilde{s}(n-1) \tag{1}$$

- Empaquetado de marcos: La salida de la pre énfasis es empaquetada en marcos de N muestras, con marcos adyacentes los cuales son separados en muestras M. Si $x_i(n)$ es el l^{th} marco del habla, y hay L marcos con señal del habla entera, entonces

$$x_i(n) = \tilde{s}(Ml + n) \tag{2}$$

  donde $(n = 0, 1, ..., N)$ y $(l = 0, 1, ..., L-1)$
- Ventaneo: Después de empaquetar en marcos, el siguiente paso es que a cada marco se le minimizan las discontinuidades de la señal de principio a fin. Si definimos la ventana como $w(n), 0 \leq n \leq N-1$ entonces el resultado del ventaneo es la señal:

$$\tilde{x}(n) = x_i(n)w(n) \tag{3}$$

  donde $0 \leq n \leq N-1$

- Analisis de autocorrección: El siguiente paso es correlacionar cada marco de señal ventaneada en orden para dar

$$r_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n)\tilde{x}_l(n+m) \tag{4}$$

donde el valor de autocorrección más alto, (p), es el orden del análisis CPL.

- Análisis CPL: El siguiente paso es el análisis CPL, donde se convierte cada marco de $(p+1)$ autocorrecciones a un conjunto de parámetros CPL usando el método de Durbin. Esto puede ser dado mediante el siguiente algoritmo:

$$E^{(0)} = r(0) \tag{5}$$

$$k_i = \frac{r(i) - \sum_{j=1}^{i-1} \alpha_j^{i-1} r(|i-j|)}{E^{i-1}} \tag{6}$$

$$\alpha_j^{(i)} = k_i \tag{7}$$

$$\alpha_j^{(i)} = \alpha_j^{(l-i)} - k_i \alpha_{i-j}^{(i-1)} \tag{8}$$

$$E^{(i)} = (l - k_i^2)E^{i-1} \tag{9}$$

Al resolver de 5 a 9 recursivamente para $i = 1, 2, ..., p$, el coeficiente CPL, a_m, es dado como

$$a_m = \alpha_m^{(p)} \tag{10}$$

- Conversión de parámetros CPL a coeficientes cepstrales: Los coeficientes cepstrales pueden ser derivados directamente del conjunto de coeficientes CPL. La recursión usada es
Para $1 \leq m \leq p$

$$c_m = a_m + \sum_{k=1}^{m_1} (\frac{k}{m}) * c_k * a_{m-k} \tag{11}$$

Para $m \geq p$

$$c_m = \sum_{k=m-p}^{m-1} (\frac{k}{m}) * c_k * a_{m-k} \tag{12}$$

3.2. Modelo oculto de Markov (MOM)

Es un modelo estadístico donde se asume que el sistema a modelar es un proceso de Markov de parámetros desconocidos. El objetivo es determinar los parámetros desconocidos de una cadena a partir de los parámetros observables. Los parámetros extraídos se pueden emplear para llevar a cabo análisis sucesivos. En un modelo oculto de Markov, el estado no es visible directamente, lo son las variables influidas por el estado. Cada estado tiene una distribución de probabilidad sobre los posibles símbolos de salida. Consecuentemente, la secuencia de símbolos generada por un MOM proporciona cierta información acerca de la secuencia de estados. Los modelos ocultos de Markov son aplicados a reconocimiento de formas temporales, como reconocimiento del habla, de escritura manual, de gestos, etiquetado gramatical o en bioinformática. En el reconocimiento de voz se emplea para modelar una frase completa, una palabra, un fonema o trifonema en el modelo acústico.

La Figura 1 muestra la arquitectura general de un MOM. Cada óvalo representa una variable aleatoria que puede tomar determinados valores. La variable aleatoria $x(t)$ es el valor de la variable oculta en el instante de tiempo t. La variable aleatoria $y(t)$ es el valor de la variable observada en el mismo instante de tiempo t., las flechas indican dependencias condicionales. El valor de la variable oculta $x(t)$ (en el instante t) solo depende del valor de la variable oculta $x(t-1)$ (en el instante $t-1$). A esto se le llama propiedad de Markov.

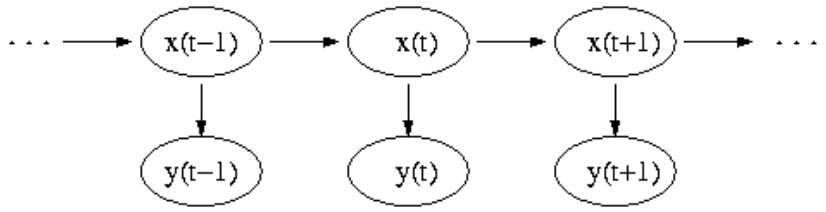

Fig. 1. Diagrama de la arquitectura general de un MOM.

Representación formal del modelo oculto de Markov
Una notación común del MOM es representarlo como una tupla:

$$(Q, V, \pi, A, B)$$

donde:

- El conjunto de estados $Q = 1, 2, ..., N$
  - El estado inicial se denota como q_t
  - En el caso de la etiquetación, cada valor de t hace referencia a la posición de la palabra en la oración.
- El conjunto V representa los posibles valores $_1, v_2, ..., v_M$ observables en cada estado

  - M es el número de palabras posibles y cada v_k hace referencia a una palablra diferente.
- $\pi = \pi_i$ son las probabilidades iniciales, donde:
  - π_i es la probabilidad de que el primer estado sea el estado Q_i
- El conjunto de probabilidades de transiciones entre estados se denota por $A = a_{ij}$

$$a_{ij} = P(q_t = j | q_{t-1=i}) \tag{13}$$

  donde, a_{ij} es la probabilidad de estar en el estado j es el instante t si en el instante anterior $t-1$ estaba en el instante i.
- El conjunto de probabilidades de las observaciones se representa por $B = b_j(v_k)$.
  $b_j(v_k) = P(o_t = v_k | q_t = j)$, es decir, la probabilidad de observar v_k cuando se está en el estado j en el instante t.
- La secuencia de observables se denota como un conjunto $O = (o_q, o_2, ..., o_T)$.

Los Modelos ocultos de Markov han demostrado ser una técnica efectiva en el procesamiento del Reconocimiento Automático del Habla. Para este trabajo se aplicará dicha técnica en el Modelo Acústico donde servirá de ayuda para la extracción de formantes de palabras, fonemas, o incluso de frases completas.

3.3. Coeficientes cepstrales en frecuencia MEL

Una técnica de extracción de parámetros de las más importantes y utilizadas actualmente en varios sistemas de reconocimiento de voz, es la obtención de los coeficientes de frecuencia Mel (CFM). Los coeficientes CFM son un tipo particular de coeficientes cepstrales derivados de la aplicación del Cepstrum sobre una ventana de tiempo de la señal de voz. El concepto de coeficientes CFM surge de hacer uso de una nueva escala de frecuencia no lineal denominada MEL para imitar el comportamiento psicoacústico a tonos puros de distinta frecuencia dentro del oído humano. De hecho, estudios dentro de esta ciencia han demostrado que el sistema auditivo humano procesa la señal de voz en el dominio espectral, caracterizándose por tener mayores resoluciones en bajas frecuencias y esto es precisamente lo que se consigue mediante la escala MEL, asignar mayor relevancia a las bajas frecuencias de forma análoga a como se hace en el sistema auditivo humano, en concreto en el oído interno. La obtención de los coeficientes MFCC ha sido considerada como una de las técnicas de parametrización de la voz más importante y utilizada dentro del área de verificación de interlocutor. El objetivo de esta transformación es obtener una representación compacta, robusta y apropiada para posteriormente poder obtener un modelo estadístico del locutor con un alto grado de precisión. Para obtener los coeficientes cepstrales en frecuencia MEL se aplica la Ecuación 14.

$$C_{MFCC}[m] = \sum_{k=0}^{N-1} log(E_k) cos\left(m\left(d - \frac{1}{2}\right)\frac{\pi}{N}\right) \tag{14}$$

donde:

- $m = m - esimo$ coeficiente MEL calculado.
- d = número de filtros utilizados en el banco de filtros MEL
- N = Tamaño de la Transformada Discreta de Fourier aplicada a la señal de voz enventanada.
- E_k = Energía correspondiente a cada uno de los F filtros

Particularmente, consideramos que ésta forma de parametrización de la señal de voz es muy conveniente y fácil de obtener. Sustentándonos en la teoría presentada, los coeficientes cepstrales en frecuencia MEL son parámetros que ofrecen información relevante de una señal de voz, además que permiten separar las dos componentes de información de la misma: la entonación y del tracto vocal.

4. Corpus de reconocimiento automático del habla

Los principales corpus a utilizar dentro de la propuesta multilenguaje, son mencionados a continuación:

- Corpus PRESEEA [17] el cual tiene como principal objetivo identificar los rasgos característicos del español hablado de Valencia. Este nace en 1996 por el equipo de investigación PRESEEA, coordinado por el Dr. José Ramón Gómez Molina. Las muestras recopiladas corresponden a 72 entrevistas semidirigidas con informantes de 3 niveles socioculturales y con un contenido aproximado de 425.000 palabras. Dicho Corpus, facilita la identificación de los rasgos característicos del castellano usado por los hablantes de dicha área metropolitana en un registro comunicativo semiformal o neutro.
- Corpus de Inglés de la Universidad de Edimburgo (EUSTACE). El Corpus EUSTACE [14] comprende 4608 oraciones habladas grabadas en el departamento de Lingüística Teórica Aplicadas de la Universidad de Edimburgo. Estas oraciones son mencionadas por seis hablantes del inglés británico, 3 mujeres y 3 hombres y fueron diseñadas para examinar el número de efectos duracionales en la voz y están controladas por su longitud y contenido fonético. En la Figura 2 se muestra la señal de voz y el espectrograma de una muestra de voz perteneciente al corpus EUSTACE.
- DIMEx100 y DIME (Diálogos Inteligentes Multimodales en Español). El Corpus DIMEx100 [15] tiene por objetivo hacer posible la construcción de modelos acústicos y diccionarios de pronunciación para la creación de sistemas computacionales para el reconocimiento del español hablado en México. Este tipo de sistemas permiten transcribir una señal de voz en su representación textual.

4.1. Tabla comparativa de los corpus utilizados

En la Tabla 1 que contiene los elementos principales de los tres corpus a utilizar en este trabajo, tomando como características principales el número de muestras, número de locutores que interfieren y el tipo de muestra.

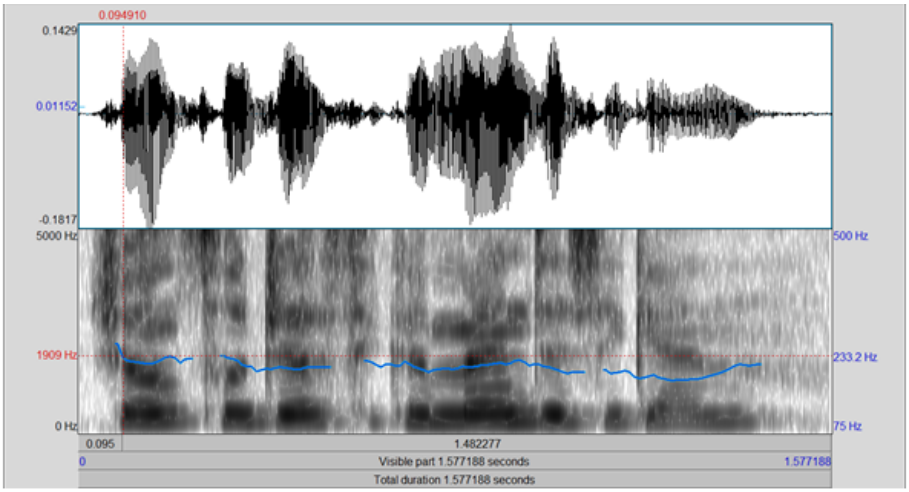

Fig. 2. Espectrograma de voz de la frase "John saw Jessica mend it again"pronunciada por una mujer en Inglés Británico.

Tabla 1. Tabla comparativa de corpus utilizados.

Nombre del Corpus	No. de muestras	No. de locutores	Tipo de muestra
PRESEEA	72 Entrevistas	4	Muestreo con extensión fija y exhaustiva.
EUSTACE	4608 oraciones	6	Formato ESPS y WAV, a una tasa de muestreo de 16 KHz y 24 dB de magnitud.
DIMEx100 y DIME	5010 oraciones	100	Formato mono a 16 bits y a 44.1 kHz, bajo *Wave Label*.

5. Sistema de reconocimiento automático del habla para corpus multilenguaje del locutor (RAHM)

La Figura 3 muestra cada uno de los corpus que sirven como entrada a ésta propuesta, los cuales llevan por nombre PRESEEA, EUSTACE y DIMEx100 respectivamente, previamente descritos. Cada uno de los componentes integrados en cada corpus deberán pasar por una extracción de características, donde podremos examinar más a fondo cada una de las partes resultantes de los corpus. Se obtendrán entonces ciertos formantes resultantes de cada corpus cada uno bajo diferentes métodos, es decir, MOM, CFM y CPL respectivamente para PRESEEA, EUSTACE y DIMEx100. La propuesta de Reconocimiento del Habla Mutilenguaje en la etapa de extracción de características, se puede apreciar en la Figura 3.

A continuación se muestra el avance de dicha propuesta, donde se ha realizado

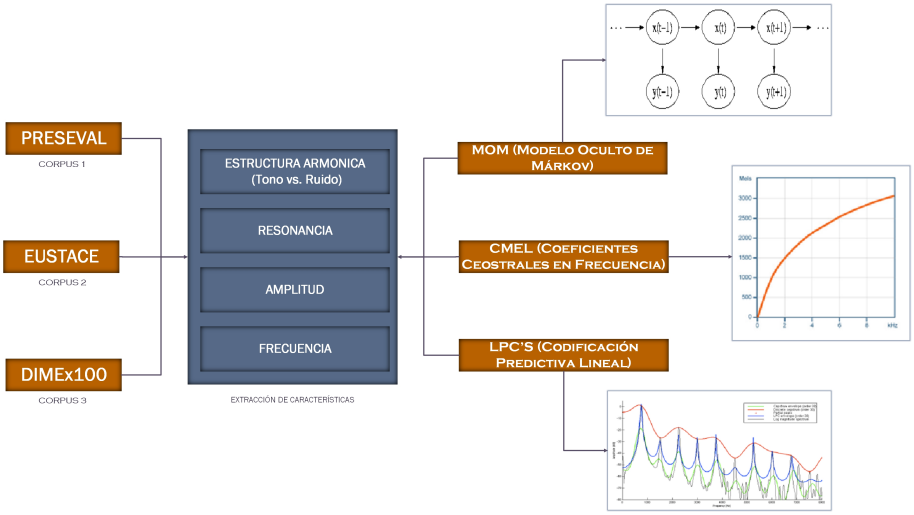

Fig. 3. Propuesta para el sistema de reconocimiento multilenguaje bajo MOM, CMEL y LPC's, en la etapa de extracción de formantes.

la extracción de las caracter—'isticas de los 3 corpus de voz anteriormente descritos.

5.1. Obtención de las características del corpus PRESEEA

El uso de la herramienta de PRAAT nos sirve para analizar u obtener características de audio, las cuales en ocasiones se obtienen a través de una señal o una voz, para posteriormente darle un uso específico dependiendo de nuestras necesidades. El objetivo es analizar el Corpus "PRESEEA" para posteriormente obtener la extracción de características que ayudará a la identificación del hablante. A continuación se enlistaran los pasos para la extracción de características de las muestras: (a) se seleccionan las 10 muestras desde la ventana Praat Objects, se visualiza el espectograma y se determina el rango de 3500Hz, (b) se obtiene el Pitch (Tono, Hz) y se obtiene la intensidad (db), (c) se hace el cálculo de sus formantes, de modo que se pueda visualizar su trayectoria a lo largo de la onda. En la Figura 4 se muestra la matriz de las 10 muestras de voz que fueron obtenidas por el Corpus PRESEEA y donde se muestra los valores correspondientes a las características obtenidas.

5.2. Obtención de las características del corpus DIMEx100

Para obtener las características de voz de dicho corpus fue necesario separar las palabras por cada oración dicha por el locutor, obtenidas por medio del programa Praat. Dentro del corpus se encuentran un total de 7 locutores diferentes donde intervienen 2 mujeres y 5 hombres. A cada palabra se le hizo la extracción

CORPUS PRESEEA

Muestras	Punto en el espectrograma	Segundos	(Análisis del tono Hz)	(Análisis de la intensidad dB)	Formantes			
Conv1	3500	3.609011	109.986399	78.931595	269.607769	1305.65853	2224.46005	3677.115703
Conv2	3500	1.128443	98.642629	86.079772	260.655081	1699.79315	2654.19012	3473.443557
Conv3	3500	1.003868	105.133992	82.792438	342.239126	1693.29933	2661.80493	3521.98262
Conv4	3500	1.203868	107.31849	77.291366	304.515062	1694.41563	2497.10435	3503.582736
Conv5	3500	1.375867	90.576321	89.471606	335.009811	1573.32754	2731.91205	3518.874236
Conv6	3500	0.83764	116.714949	82.724015	340.700786	1256.72135	2332.20479	3430.623548
Conv7	3500	0.157067	116.778331	81.547988	302.456841	1611.80108	2645.7048	3492.397091
Conv8	3500	0.132583	113.007869	81.794636	298.632725	1590.04476	2632.58269	3467.177862
Conv9	3500	1.715219	120.404686	84.052438	362.239126	1513.29933	2171.80493	35423.98262
Conv10	3500	1.219037	118.678473	85.026402	332.28554	1305.70486	2145.18913	3133.558223

Fig. 4. Tabla de características calculadas para el corpus PRESEEA.

del tono el cual es medido en hertz y así mismo la intensidad, la cual esta dada en decibeles, y por último se hizo la extracción de los 4 formantes, medidos en hertz. Aunque cada muestra varió en tiempos, se estableció una frecuencia de 3500 Hertz como condición inicial como en el corpus anterior. Para seguir con el procedimiento, fue necesario conocer por cada locutor el promedio del tono, intensidad y los formantes. Después de obtener cada una de las características para cada muestra se hizo el registro en una tabla y/o matriz de la cual en la Figura 5 se muestran los diez ejemplos tomados para realizar las gráficas correspondientes de cada una de las características como el tono, la intensidad y los formantes.

Muestra	Pitch (Hz)	Intensidad (dB)	Formante 1	Formante 2	Formante 3	Formante 4
locutor1_cual	166.771319	83.586892	523.202906	821.64902	2532.32265	3436.013358
locutor2fem_avancemos	286.335852	81.989834	577.37283	1954.711557	2909.773112	4267.363891
locutor3mas_estamos	164.312667	62.67601	491.08748	1869.265095	2579.405397	3975.68134
locutor3mas_puntos	164.412107	65.090599	433.702412	934.295444	2405.740231	3761.202684
locutor4fem_departamento	191.660023	70.958987	636.050739	1632.534055	2595.586095	4103.359404
Locutor5_explicar	155.082241	76.158301	394.5892706	2216.978903	3481.371294	3675.838238
Locutor6_41_herramienta	143.147655	71.41522	519.439371	1871.973696	2573.135455	3750.49384
Locutor7_31_competencia	90.656473	69.578603	430.189729	1487.581143	2317.03755	3576.601287
Locutor6_32_posible	148.743834	75.112057	404.245684	2113.509593	3460.688218	4070.694819
Locutor5_43_desarrollado	92.228238	76.894352	487.025495	1679.91227	2940.942573	3443.590818

Fig. 5. Matriz de las características obtenidas del corpus DIMEx100.

5.3. Obtención de las características de (EUSTACE)

El corpus de voz cuenta con 4608 oraciones y 6 locutores: 3 hombres y 3 mujeres. Debido a que el tamaño del corpus es excesivamente grande, sólo se ha tomado una pequeña parte para la obtención de características. La porción tomada involucra 50 frases mencionadas por cada locutor. De cada una de esas frases se obtuvieron las siguientes características: Análisis de Tono, Análisis de Intensidad y Análisis de Formantes. Dichas características fueron tomadas a un nivel de frecuencia estándar de 3500 Hz. Cada uno de los archivos de audio contiene alrededor de 15 frases, por lo que para efectuar un análisis fue necesario tomar sólo la señal comprendida por cada frase, lo que implica tomar en cuenta el instante en el que se tomó la muestra. La matriz de características se compone de 300 señales analizadas, 50 por cada locutor, y 8 valores característicos relacionados con los puntos anteriormente mencionados (Frecuencia (Hz), Tiempo (s), Tono (Hz), Intensidad (dB), Formante 1-4 (Hz)). En la Figura 6 se presenta un extracto de las primeras 10 muestras con sus respectivas características.

Características del Corpus de Voz "EUSTACE"

Muestra	Frec (Hz)	T(s)	Pitch (Hz)	Intensidad (dB)	Form_1 (Hz)	Form_2 (Hz)	Form_3 (Hz)	Form_4 (Hz)
Locutor 1 Masculino Grupo 1								
m1/capae_1	3500	1.019	120.80644	68.610449	470.629965	1154.53321	3074.46535	3696.3701
m1/capae_2	3500	4.323	128.72135	68.178712	458.268372	1204.96584	2080.47514	3885.94491
m1/capae_3	3500	7.561	122.66044	71.33897	421.043548	1105.42664	2206.74614	3354.31847
m1/capae_4	3500	11.45	139.80837	69.593722	421.02079	1454.85417	2234.0941	3881.95069
m1/capae_5	3500	15.82	126.64631	71.538076	459.873574	1211.60622	2024.20717	3445.60513
m1/capae_6	3500	20.08	125.68676	71.040469	436.916172	1180.34927	2290.04377	3573.48219
m1/capae_7	3500	23.76	121.88977	68.320104	443.776884	1389.25531	2027.13849	3585.39417
m1/capae_8	3500	27.64	129.33968	72.183424	495.92082	1220.31347	2058.79925	3640.42604
m1/capae_9	3500	30.45	136.42908	66.342916	380.507625	1815.16089	2458.08092	3606.50476
m1/capae_10	3500	35.23	137.15905	68.816431	425.159286	1666.58828	2393.36433	3454.65783

Fig. 6. Matriz de muestras obtenidas del corpus EUSTACE.

A continuación se muestran las características extraídas de 4 formantes para cada corpus analizado. En la Figura 7 se observa la extracción de los formantes de PRESEEA, en la Figura 8 se muestran los formantes extraídos de EUSTACE, y en la Figura 9 los formantes de DIMEx100.

6. Conclusiones

El método CPL (Coeficientes de Predicción Lineal) se implementó para las muestras y se obtuvo una gráfica donde se hacía comparación de la señal original con el CPL estimado. Éste proceso contiene filtros para mejorar la señal. El algoritmo que se describió en el estado del arte también se utiliza para poder calcular el CPL a las muestras correspondientes. El corpus que se utilizo fue el

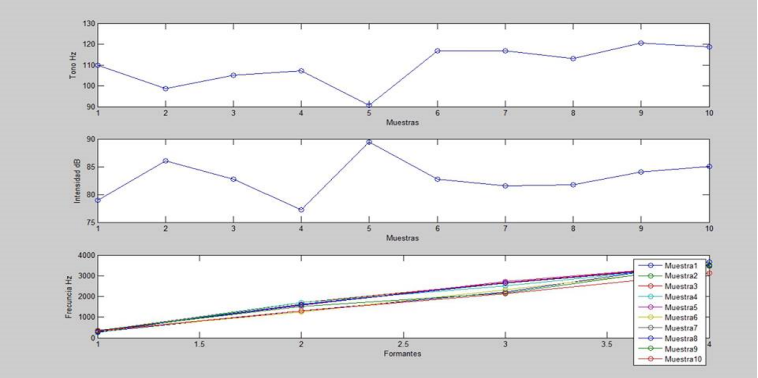

Fig. 7. Gráfica del extracción de formantes para PRESEEA.

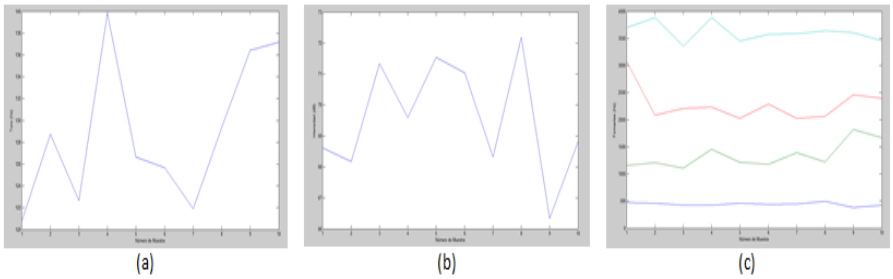

Fig. 8. Análisis de (a)Tono, (b)Intensidad y (c)Formantes de 10 frases contenidas en el corpus EUSTACE.

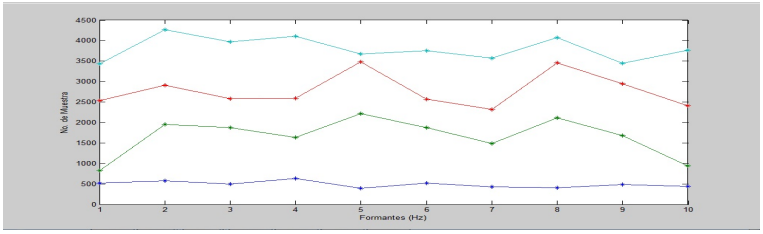

Fig. 9. Gráfica del extracción de formantes para DIMEx100.

de la Universidad de Valencia PRESEEA que se obtuvo de las respuestas de la pregunta? "Has hecho tu servicio militar?" Para las mujeres la pregunta fue "Si hubieras sido hombre / hubieras hecho el servicio militar?". De las respuestas obtenidas del hablante se extrajeron las características y se aplicó el método CPL. En el caso del Corpus DIMEx100 se encontraron diversas dificultades para su análisis. En primera, las muestras eran muy variadas en cuestión de locutores y de frases dichas por cada uno. Para su análisis fue necesario hacer el corte de las frases en palabras con el fin de lograr el reconocimiento de locutor, pero esto se dificultó al saber que las frases contenían diferentes palabras. Aun así fue relativamente sencillo obtener las características necesarias para su posterior análisis. En el caso de la técnica de los Modelos Ocultos de Markov es necesario mencionar que aunque es uno de los métodos más utilizados en el Reconocimiento Automático del Habla, era necesario adaptar nuestro corpus de diferente manera para poder hacer el procesamiento de extracción de formantes. De acuerdo al análisis que se realizó sobre la técnica antes mencionada, se puede concluir que es una de las mas eficaces para este tipo de trabajos. Particularmente, el corpus de voz EUSTACE es bastante robusto, así que sólo se consideró un 8 % aproximadamente de las señales de voz para el análisis y obtención de características. Las frases contenidas en cada archivo de audio analizado son sencillas y claras, lo que permite una fácil comparación entre las características obtenidas para cada una. El cálculo de las características fue relativamente simple, ya que son datos que se pueden obtener desde la herramienta utilizada de manera directa. Como resultado se obtuvieron 12 coeficientes en escala de mel, por cada serie de características y señal analizada, que representan aquellas frecuencias, las cuales proporcionan información relevante que puede ser útil en sistemas de Reconocimiento Automático del Habla. Es importante mencionar que la parte de la extracción de características para el resto del corpus será trabajo a futuro que resta por realizar bajo el esquema propuesto en la Figura 3.

Referencias

1. Univaso, P., Gurlekian, J. A., Evin, D.: Reconocimiento del habla continua independiente del contexto para el español de Argentina. Revista clepsidra, p. 11 (2009)
2. Grulekian, A. J., Evin, D., Torres, H., Renato, A.: Sistemas de Análisis Acústico y de Reconocimiento Automático en Habla Espontanea. Subjetividad y Procesos Cognitivos, vol. 14 (2), p. 10 (2010)
3. Solís Villarreal, J.F., Yáñez Márquez, C., Suárez Guerra, S.: Reconocimiento automático de voz emotiva con memorias asociativas Alfa-Beta SVM. Polibitis (2011)
4. Gomez, R., Tatsuya, K.: Denoising Using Optimized Wavelet Filtering for Automatic Speech Recognition. Academic Center for Computing and Media Studies (ACCMS), Kyoto University, Japan (2011)
5. Pérez, S., Pelle, P., Estienne, C., Messina, F.: Sistema de Reconocimiento de Habla en Español con adaptación al discurso. Universidad de Buenos Aires, p. 10 (2011)

6. De la Torre, A., Fohr, D., Paul, H.J.: Métodos Para Reconocimiento Robusto De Voz Adquirida En Automóviles. Universidad de Granada, Dpto. de Electrónica y Tec. Comp., España (2011)
7. Cruz-Beltrán, L., Acevedo-Mosqueda, M. A.: Reconocimiento de Voz usando Redes Neuronales Artificiales Backpropagation y Coeficientes LPC. SEPI Telecomunicaciones ESIME IPN (2011)
8. Edmons, C., Hu, S., Mandle, D.: Improvement of an Automatic Speech Recognition Toolkit (2012)
9. Thiang, Soryu Wijogo.: Speech recognition using linear predictive coding and artificial neural network for controlling movement of mobile robot. Electrical Engineearing Department, Petra Christian University, Indonesia (2011)
10. Antoniol, G., Rollo, V. F., Venturi, G.: Linear Predective Coding and Cepstrum coefficients for mining time variant information from software repositories. ACM SIGSOFT Software Engineering Notes, vol. 30 (4), pp. 1-5 (2005)
11. Makhoul, J.: Linear Prediction:A tutorial review. Proc. IEEE, pp. 561-580 (1975)
12. Colaboradores de Wikipedia, Modelos Ocultos de Markov. Wikipedia,La enciclopedia libre. http://es.wikipedia.org/wiki/Modelo_oculto_de_Markov
13. Extracción de Características. http://bibing.us.es/proyectos /abreproy/12054/fichero/MEMORIA %252F8.Cap %EDtulo+3.pdf
14. EUSTACE (Edinburgh University Speech Timing Archive and Corpus of English),CSTR (The Centre for Speech Technology Research), University of Edinburgh. http://www.cstr.ed.ac.uk/projects/eustace/index.html
15. DIMEx100 y DIME(Diálogos Inteligentes Multimodales en Español),Universidad Autónoma de México Centro de Ciencias Aplicadas y Desarrollo Tecnológico de la UNAM (CATED-UNAM). http:turing.iimas.unam.mx/ luis/DIME/DIMEx100/manualdimex100/index.html
16. Mel Frecuencial Cepstral Coeficients. http://es.wikipedia.org/wiki/MFCC
17. PRESEEA. http://www.uv.es/preseea/ppal.htm